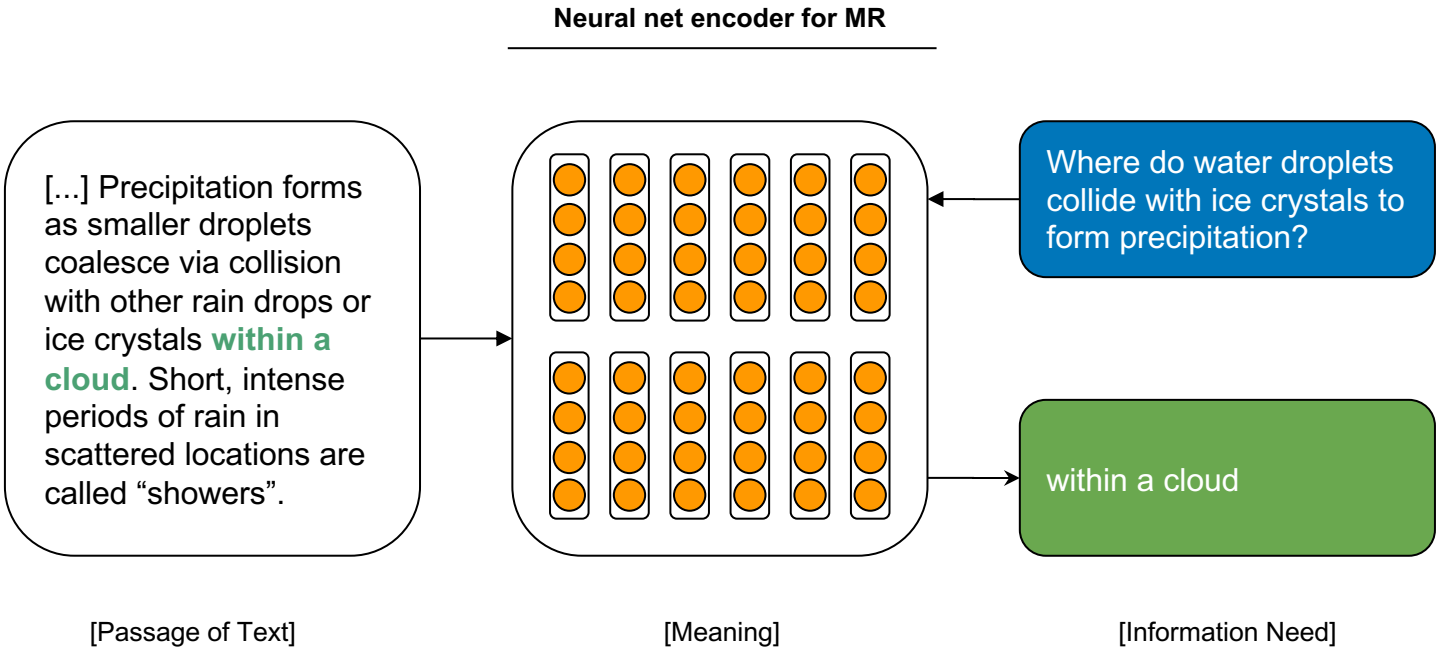
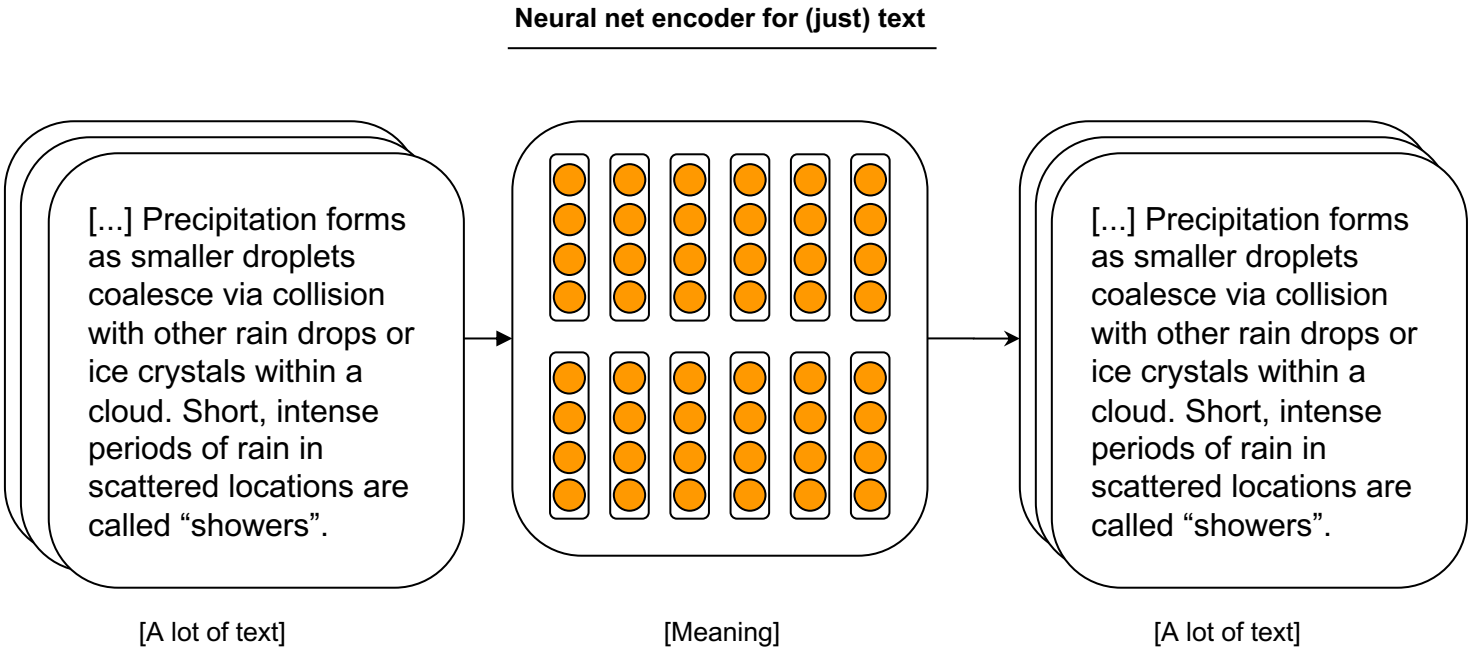
Machine Reading / Current Trend

Supervised training

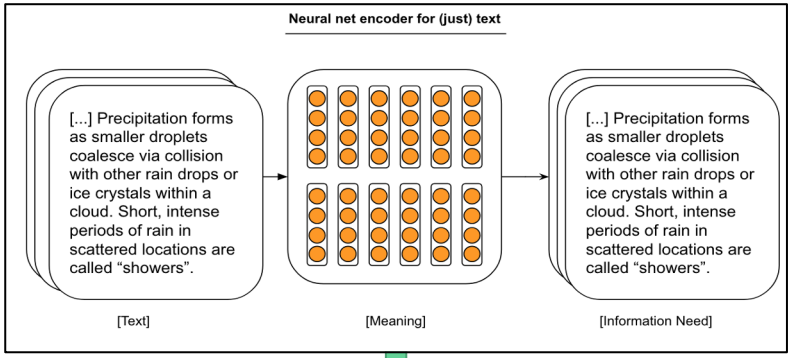


Unsupervised pretrained representations



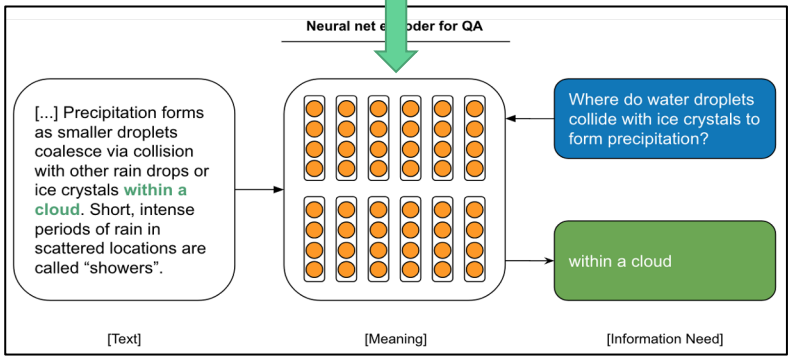
Lifting over pretrained representations

Pretrained Language Model



Transfer

Machine Reading



How is this different from pretrained word embeddings?

Pretrained Word Embeddings (word2vec)

- Predicting co-occurring of words
- Independent of other context

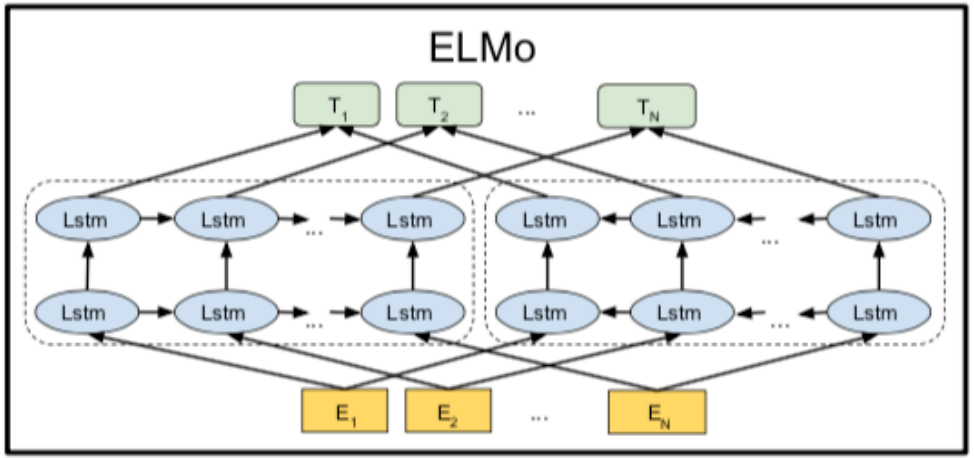
Pretrained Contextualized Embeddings (e.g. ELMo, BERT)

- Predicting whole text (using LSTM, or Self-Attention)
- Full dependence on other context

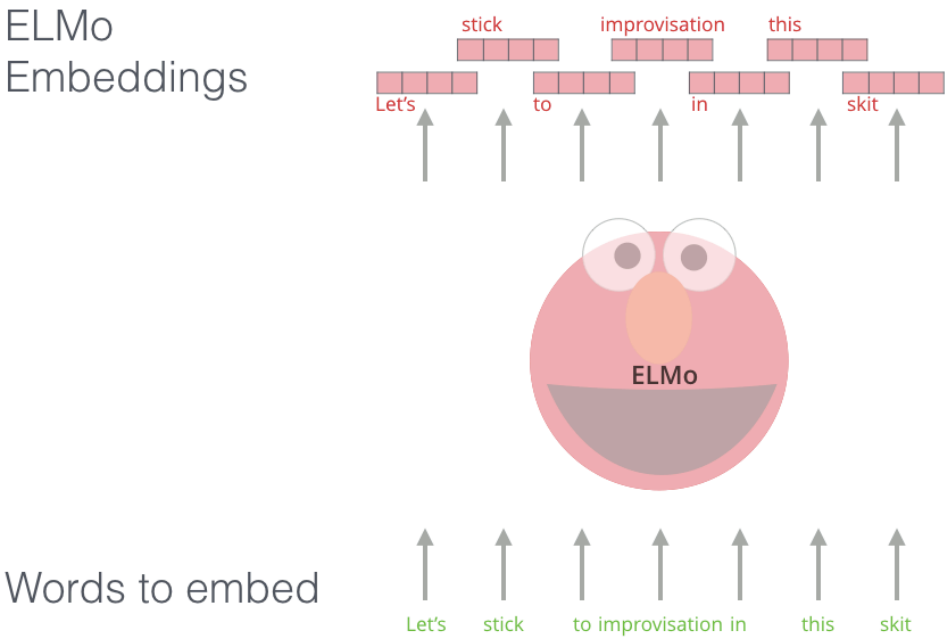
ELMo: **E**mbeddings from **L**anguage **Mo**dels

Peters et al., NAACL'18

- Train a BiLSTM for Bidirectional language modeling on a large dataset
- Run the sentence to encode through both forward and backward LSTMs
- Combine forward and backward representations into final contextual embeddings



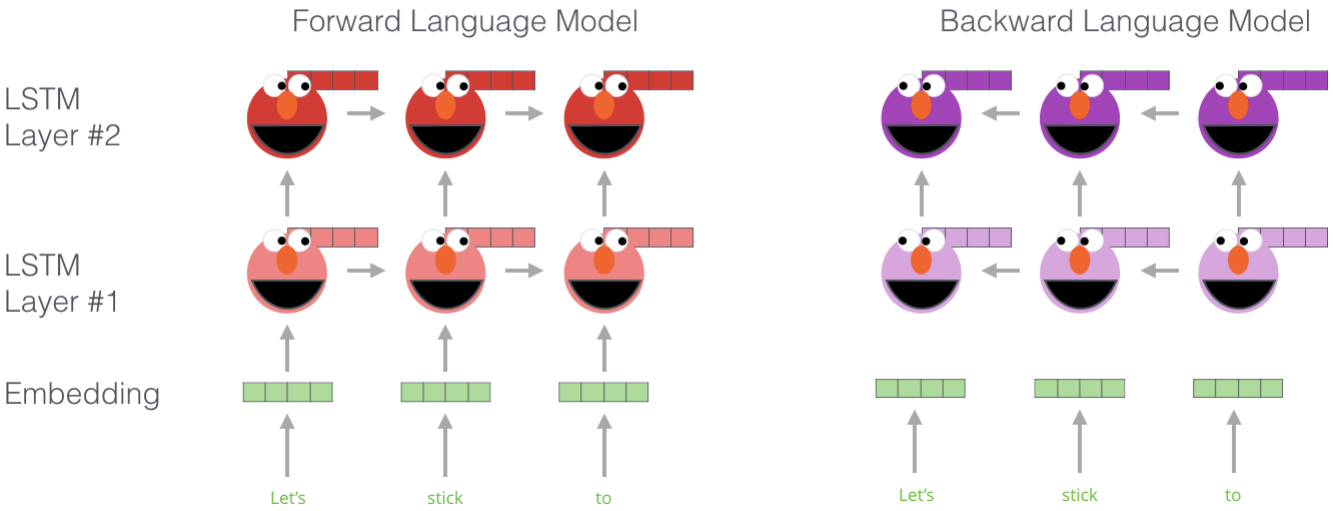
ELMo: **E**mbeddings from **L**anguage **Mo**dels



Figures from http://jalammar.github.io/illustrated-bert/

ELMo: **E**mbeddings from **L**anguage **Mo**dels

Embedding of “stick” in “Let’s stick to” - Step #1



Figures from http://jalammar.github.io/illustrated-bert/

ELMo: **E**mbeddings from **L**anguage **Mo**dels

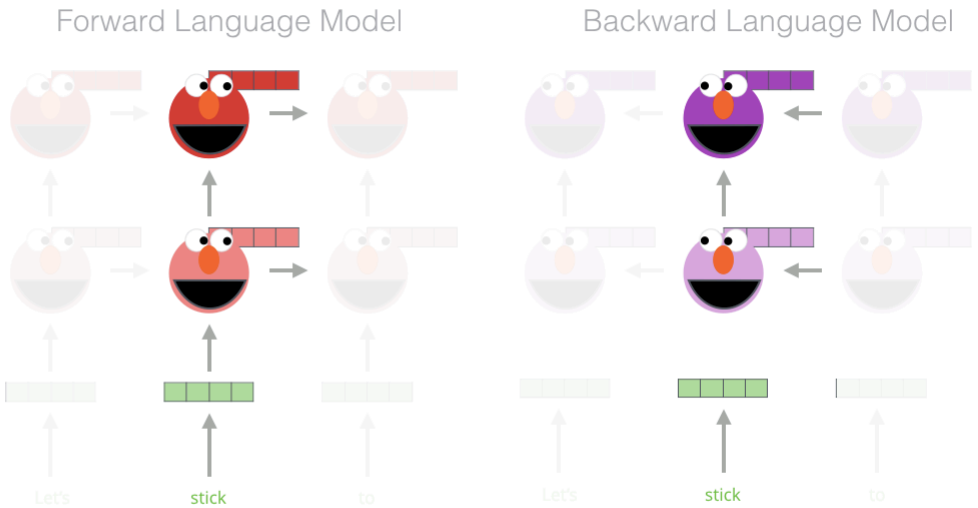
Embedding of “stick” in “Let’s stick to” - Step #2

1- Concatenate hidden layers

2- Multiply each vector by a weight based on the task

3- Sum the (now weighted) vectors

ELMo embedding of “stick” for this task in this context



Figures from http://jalammar.github.io/illustrated-bert/

ELMo performance

	Task	Previous SOTA		Our baseline	ELMo + baseline	Increase (absolute/ relative)
Machine Reading-	SQuAD	Liu et al. (2017)	84.4	81.1	85.8	4.7 / 24.9%
Textual Entailment -	SNLI	Chen et al. (2017)	88.6	88.0	88.7 ± 0.17	0.7 / 5.8%
Semantic Labeling -	SRL	He et al. (2017)	81.7	81.4	84.6	3.2 / 17.2%
Coreference Resolution -	Coref	Lee et al. (2017)	67.2	67.2	70.4	3.2 / 9.8%
Entity Extraction -	NER	Peters et al. (2017)	91.93 ± 0.19	90.15	92.22 ± 0.10	2.06 / 21%
Sentiment Analysis -	SST-5	McCann et al. (2017)	53.7	51.4	54.7 ± 0.5	3.3 / 6.8%

What is ELMo learning ?

- Meaning of words in context
  - POS, word sense, etc.

	Source	Nearest Neighbors
GloVe	play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM	Chico Ruiz made a spectacular play on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent play .
	Olivia De Havilland signed to do a Broadway play for Garson {...}	{...} they were actors who had been handed fat roles in a successful play , and had talent enough to fill the roles competently , with nice understatement .

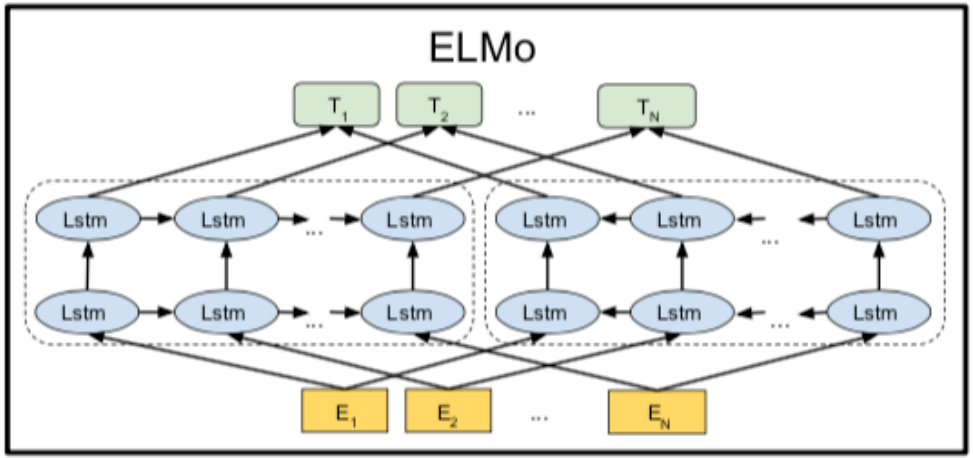
Problems with ELMo

- Need to use different architectures for different tasks
- Retraining models is slow, transfer learning is fast
- Need to deal with long term dependencies in LSTMs!

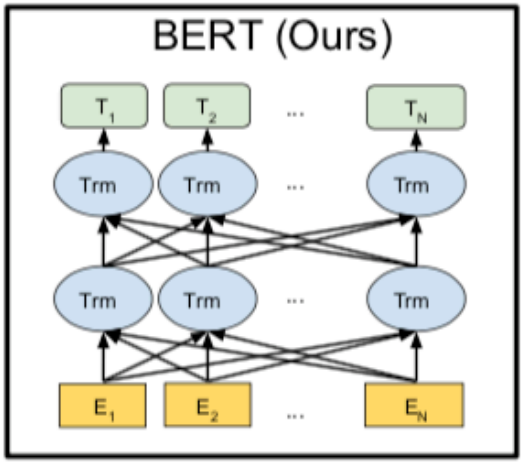
BERT - **B**idirectional **E**ncoder **R**epresentations from **T**ransformers

Devlin et al., NAACL'19

Solutions: use Transformer + encoder layers instead of decoder layers



(OpenAI GPT)

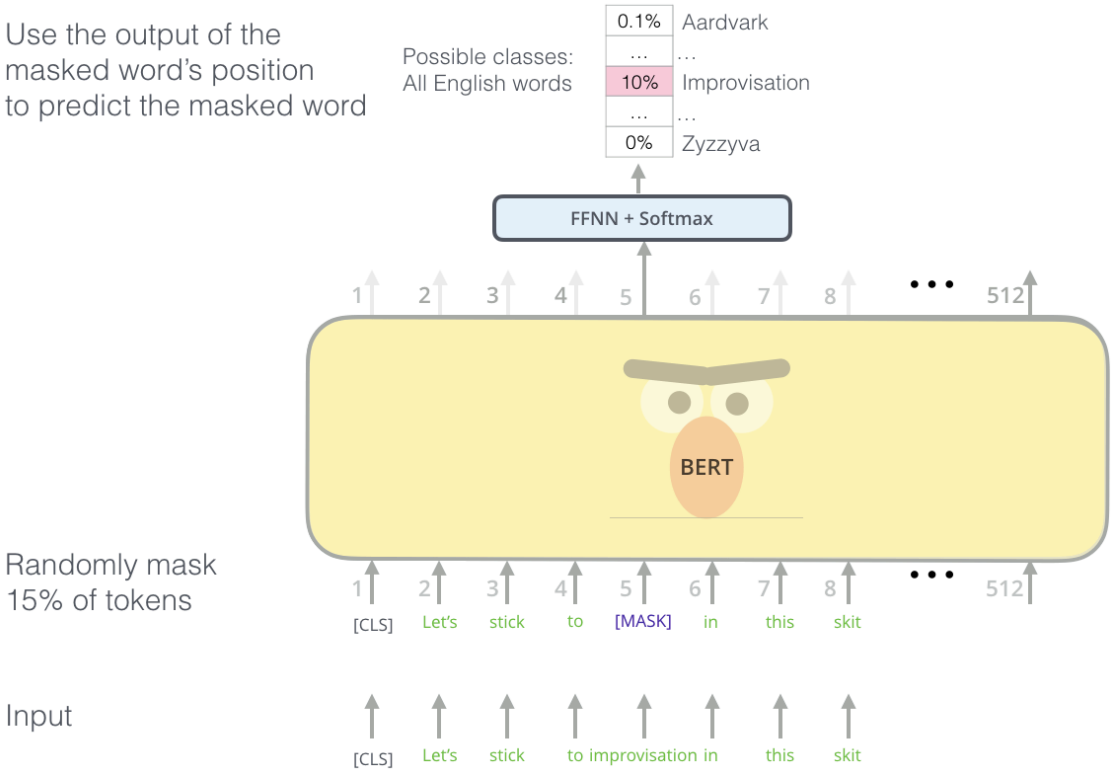


Innovation with multiple pretraining tasks

BERT – Pretraining 1: masked language modeling

- Given a sentence with some words masked at random, can we predict them?
- Randomly select 15% of tokens to be replaced with “<MASK>”

BERT – Pretraining 1: masked language modeling

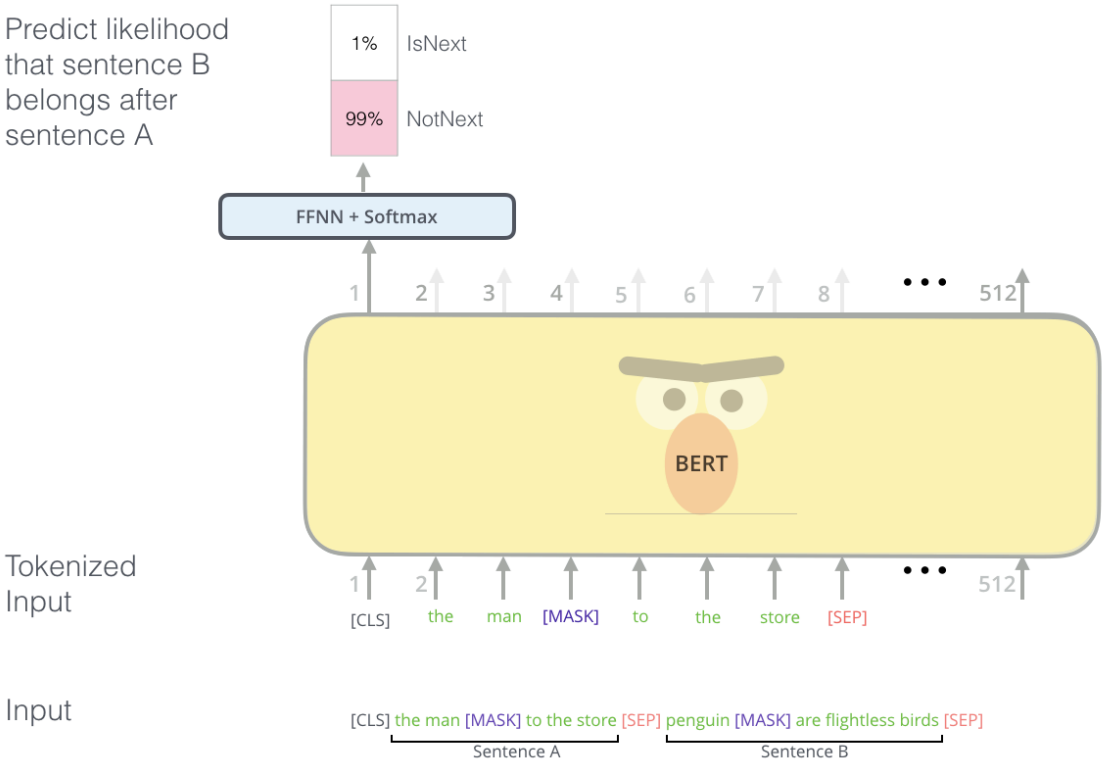


Figures from http://jalammar.github.io/illustrated-bert/

BERT – Pretraining 2: next sentence prediction

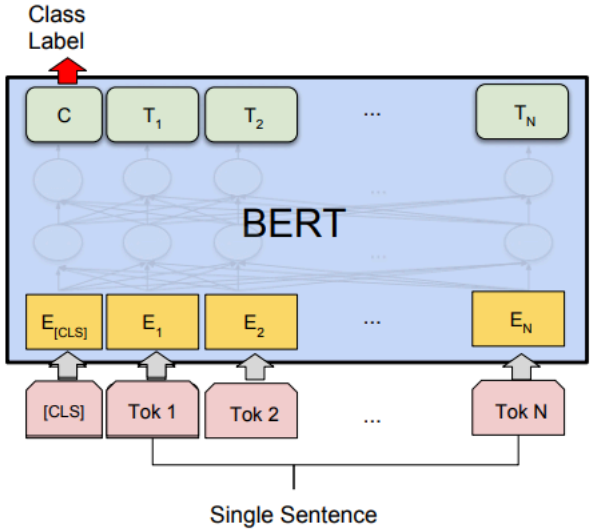
- Given two sentences, does the first follow the second?
- Teaches BERT about relationship between two sentences
- 50% of the time the actual next sentence, 50% random

BERT – Pretraining 2: next sentence prediction



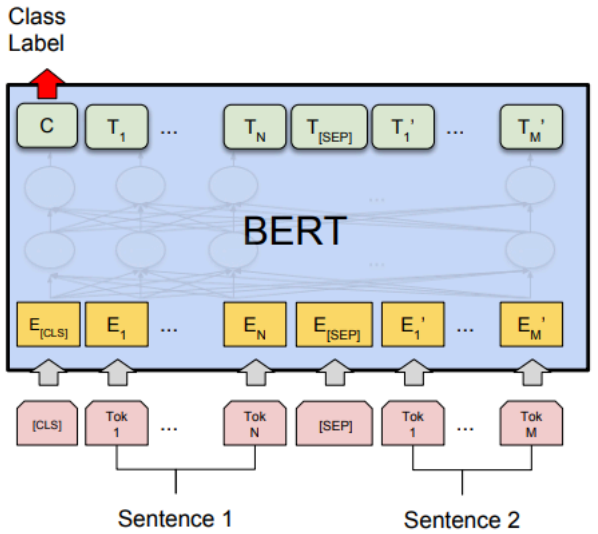
Figures from http://jalammar.github.io/illustrated-bert/

BERT – Fine-tuning for Classification



Single sentence classification

Sentiment analysis, spam detection, etc.

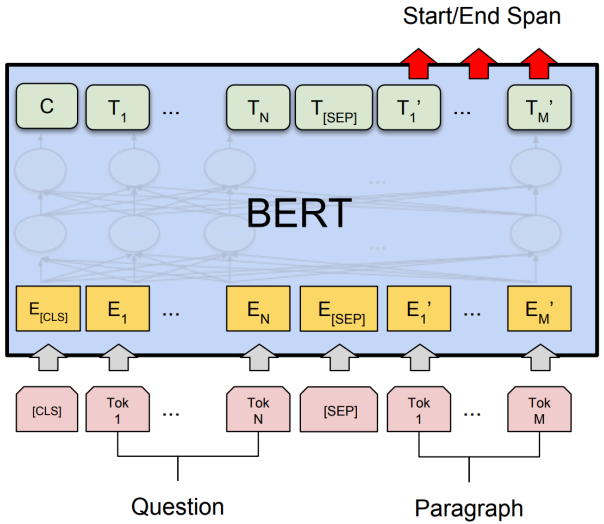


Pair of sentences classification

Entailment, paraphrase detection, etc.

Figures from Devlin et al. 18'

BERT – Fine-tuning for Machine Reading



(c) Question Answering Tasks: SQuAD v1.1

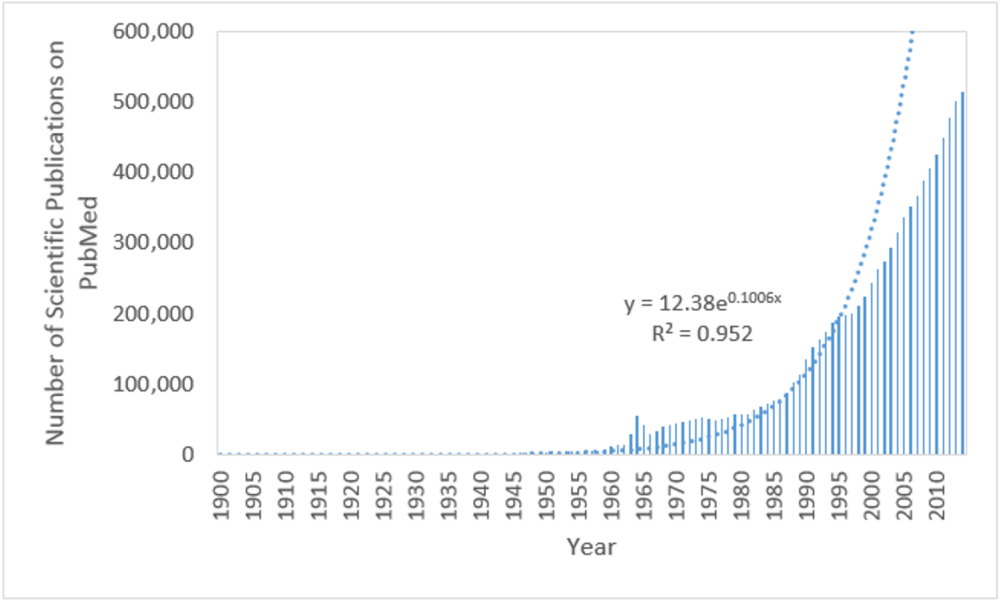
System	Dev EM	Dev F1	Test EM	Test F1
Leaderboard (Oct 8th, 2018)				
Human	-	-	82.3	91.2
#1 Ensemble - nlnet	-	-	86.0	91.7
#2 Ensemble - QANet	-	-	84.5	90.5
#1 Single - nlnet	-	-	83.5	90.1
#2 Single - QANet	-	-	82.5	89.3
Published				
BiDAF+ELMo (Single)	-	85.8	-	-
R.M. Reader (Single)	78.9	86.3	79.5	86.6
R.M. Reader (Ensemble)	81.2	87.9	82.3	88.5
Ours				
$\text{BERT}_{\text{BASE}}$ (Single)	80.8	88.5	-	-
$\text{BERT}_{\text{LARGE}}$ (Single)	84.1	90.9	-	-
$\text{BERT}_{\text{LARGE}}$ (Ensemble)	85.8	91.8	-	-
$\text{BERT}_{\text{LARGE}}$ (Sgl.+TriviaQA)	**84.2**	**91.1**	**85.1**	**91.8**
$\text{BERT}_{\text{LARGE}}$ (Ens.+TriviaQA)	**86.2**	**92.2**	**87.4**	**93.2**

Figures from Devlin et al. 18'

Open Domain Question Answering

Is Machine Reading actually useful?

Motivation 1: Information Overload

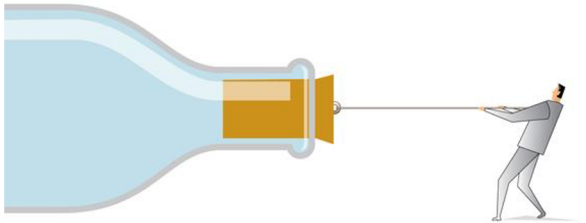


[Information Need]

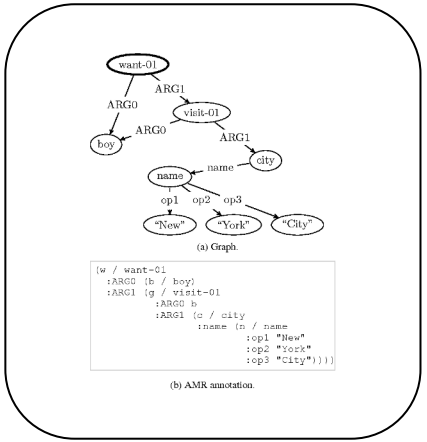
uses for

Motivation 2: The Knowledge Acquisition Bottleneck

"The problem of knowledge acquisition is the critical bottleneck problem in artificial intelligence."
E. A. Feigenbaum 1984



[Knowledge]



[Meaning]

uses for

[Information Need]

Open domain Question Answering

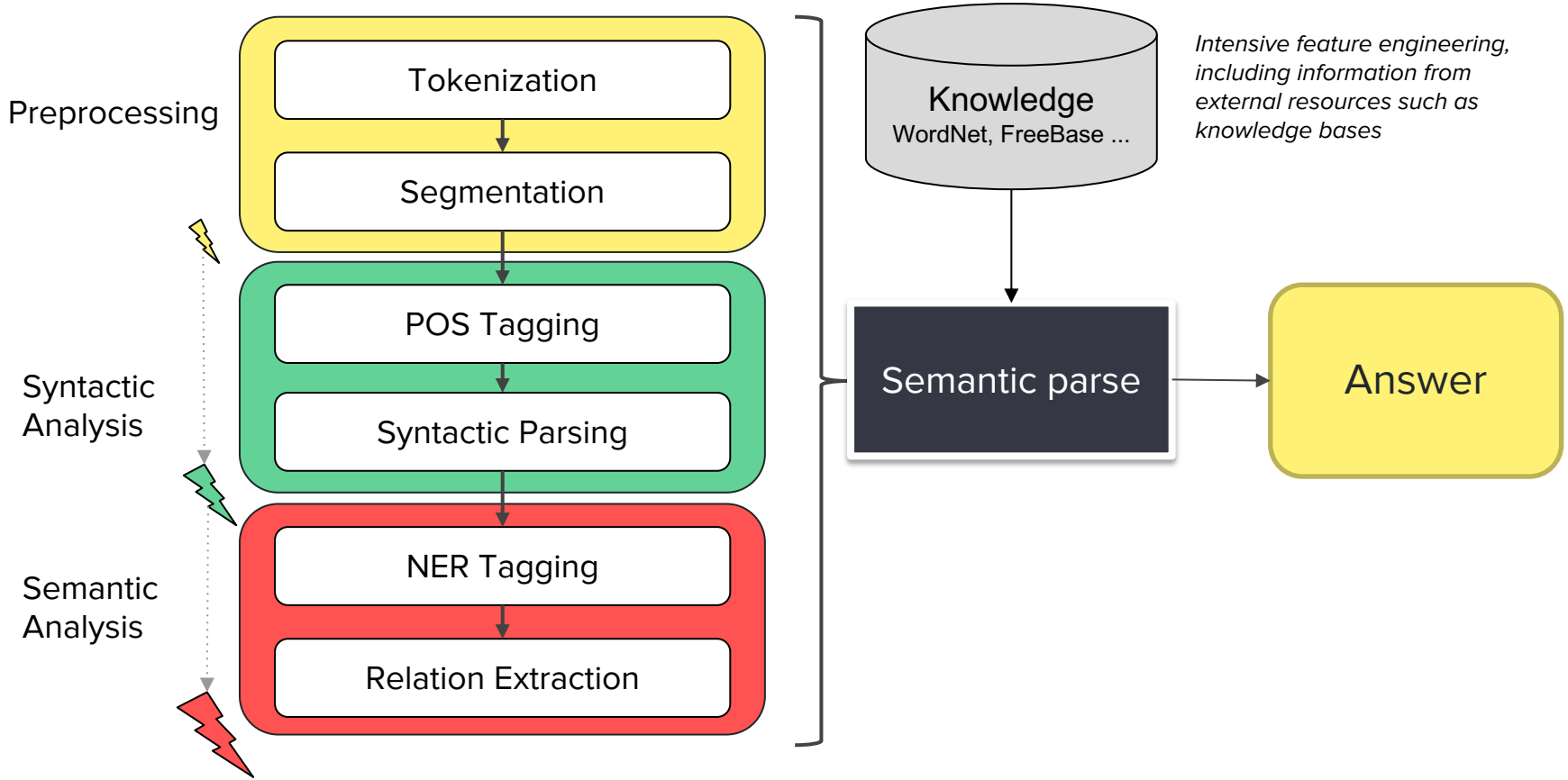
- Open domain QA: answer any question using very large knowledge sources
- Goes beyond Machines Reading that expects a paragraph to be given

- Open domain = question on any topic not a restricted subset

- In the following

1. Traditional approaches using Knowledge Bases
2. New approaches based on end-to-end Machine Reading

“Traditional” NLP for open domain QA



Semantic Parsing

Ewan forgot the mozarella in his car

[Text]

$$\exists x0\ \text{named}(x0, \text{ewan}, \text{person}) \wedge$$
$$\exists x1\ \text{mozzarella}(x1) \wedge$$
$$\exists x2\ \text{car}(x2) \wedge \text{of}(x2,x0) \wedge \text{in}(x1, x2) \wedge$$
$$\exists e\ \text{event}(e) \wedge \text{forget}(e) \wedge \text{agent}(e, x0) \wedge \text{patient}(e, x1)$$

[Meaning]

[Information Need]

Semantic parses are logical forms in PROLOG, SQL, SPARQL, etc.

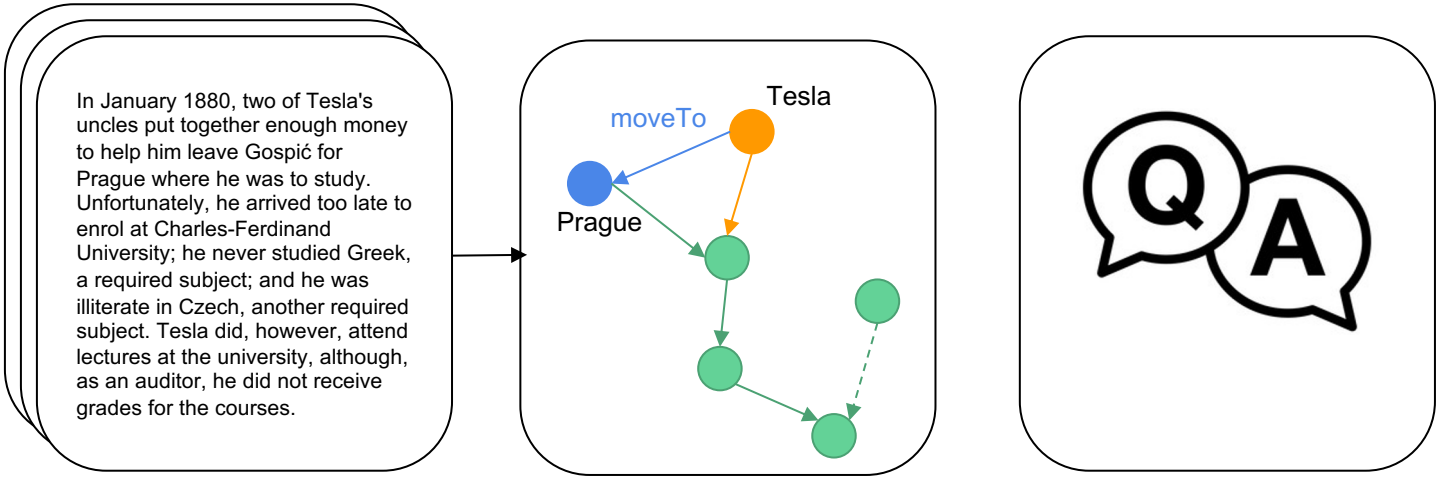
Knowledge Bases

- KB: structured repository of knowledge (usually relational DB)
- Goal: encode knowledge so that it can be queried by semantic parses **efficiently**
- Scale can be huge: billions of facts, millions of entities

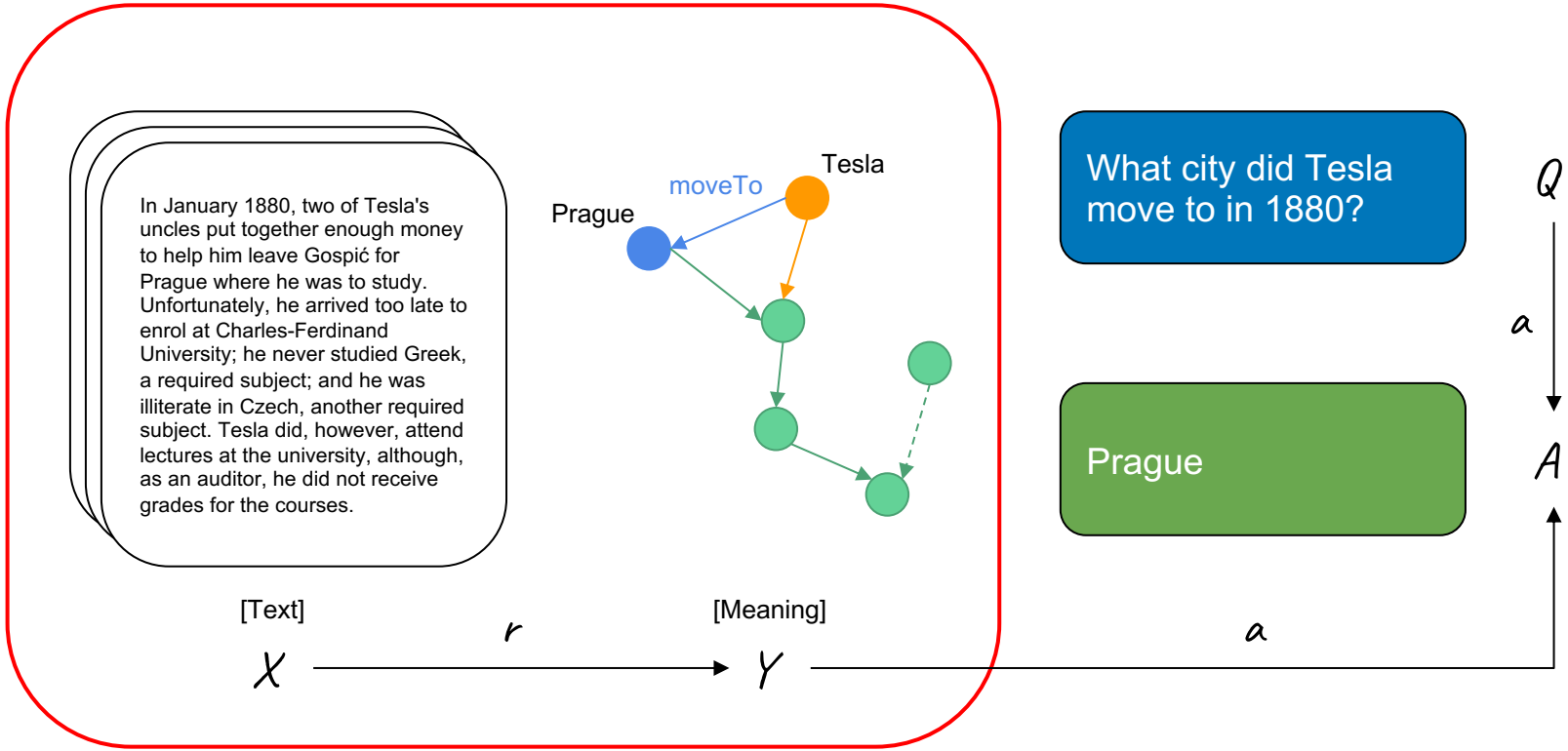
- KB can be generic or specific
- Examples: Cyc, WikiData, DBPedia, Google KG, GeneOntology, IMDB, etc.

- Key challenge is their construction!
    - Manually: Crowdsourcing, paid experts
    - Automatically: Information extraction or Automatic KB Construction

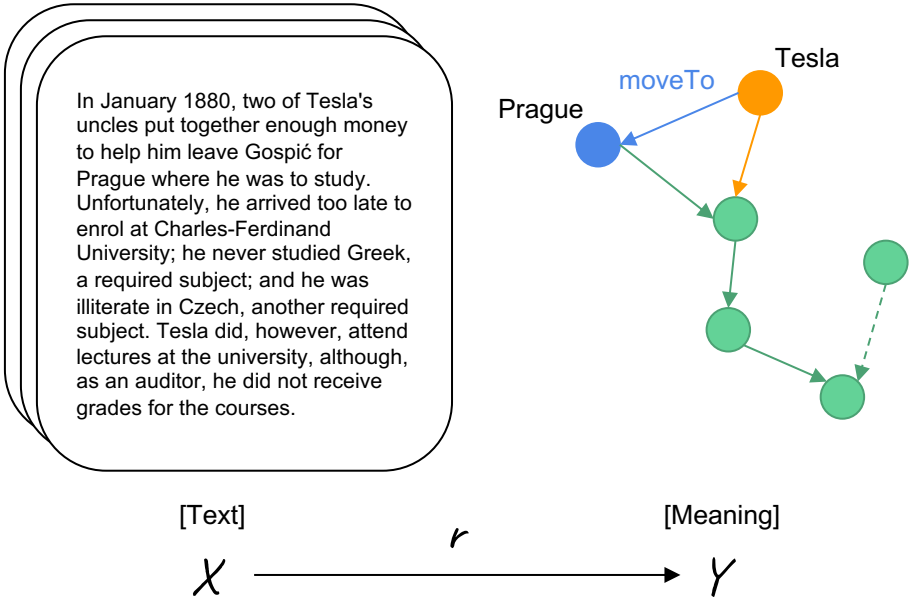
Automatic Knowledge Base Construction



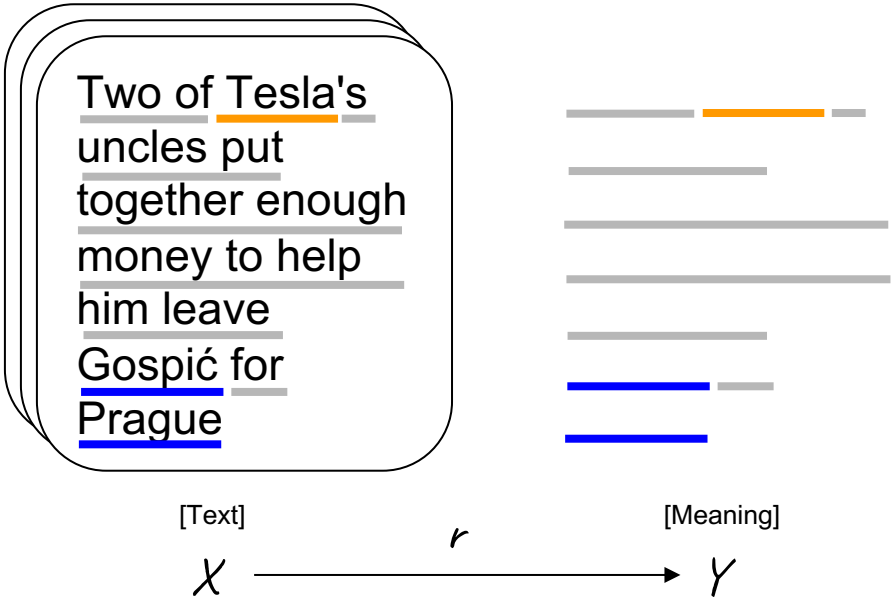
Knowledge Graph Construction



Knowledge Graph Construction



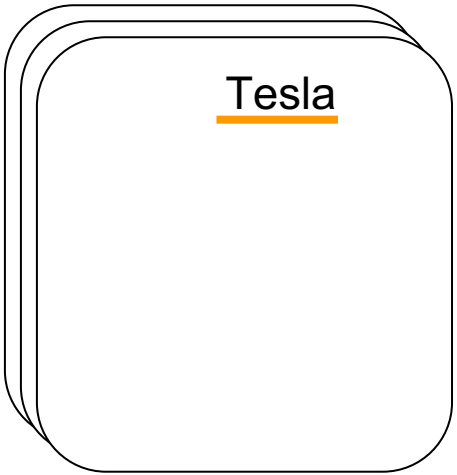
Entity Extraction



- Linear Chain CRF
- Bi-directional RNNs
- Hybrid RNN & CRFs

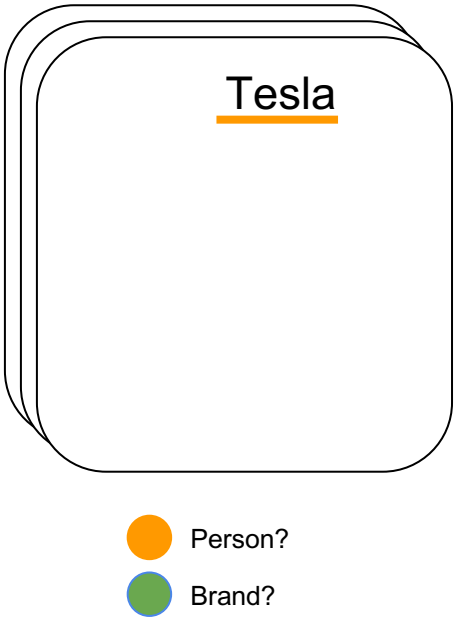
Person
Location

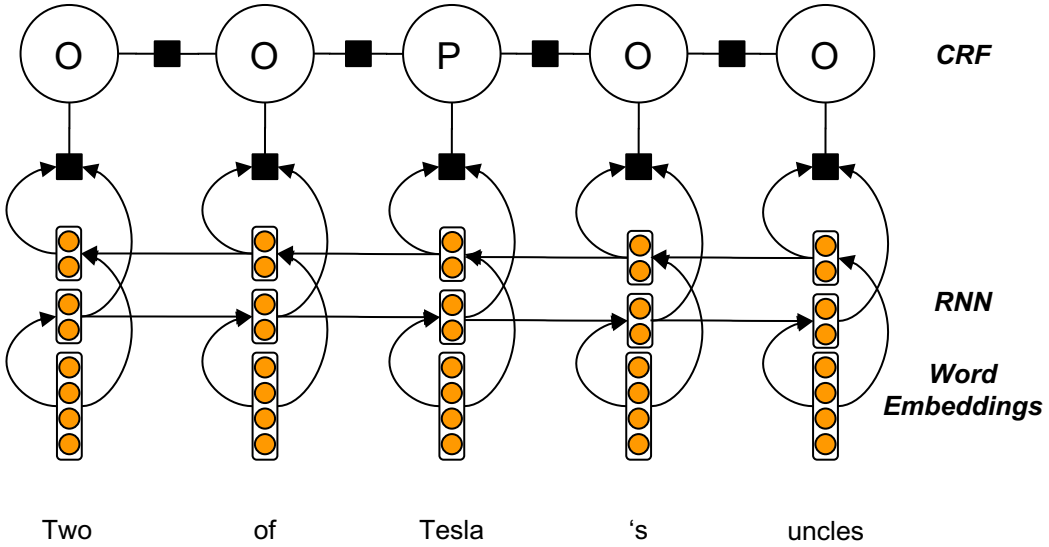
Challenge: Ambiguity



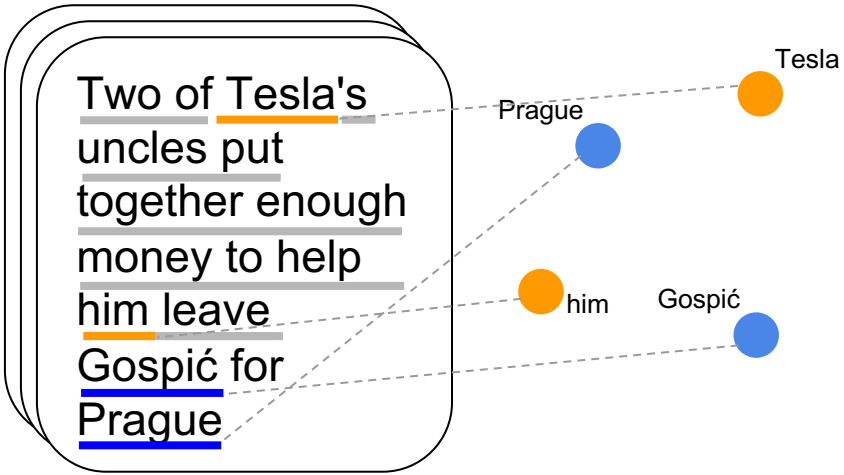
Conditional Random Fields with RNN Potentials

Huang et al., 2015

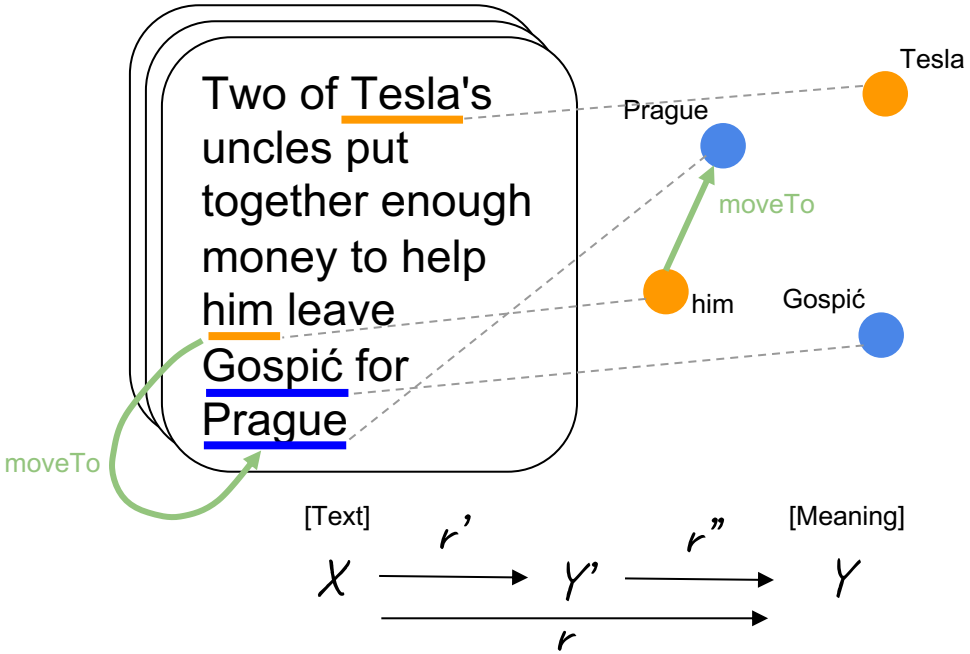




Instantiate Nodes

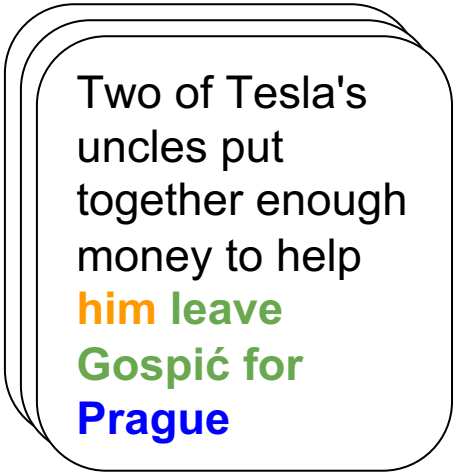


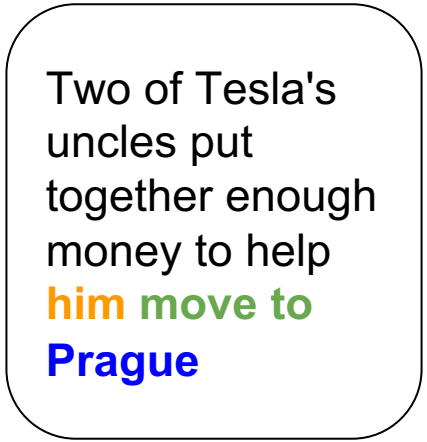
Relation Extraction

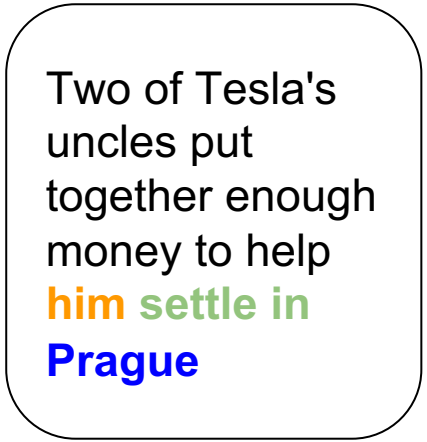


- Neural Classification
- Distant Supervision

Challenge: Variation

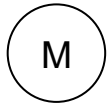






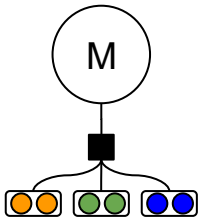
Relation Classification

[Current SOTA neural RE model]

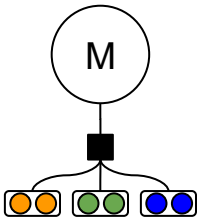


(Tesla, moveTo, Prague)

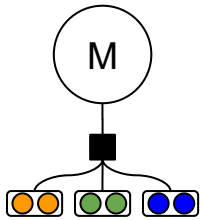
WIKIDATA



him leave Gospić for Prague



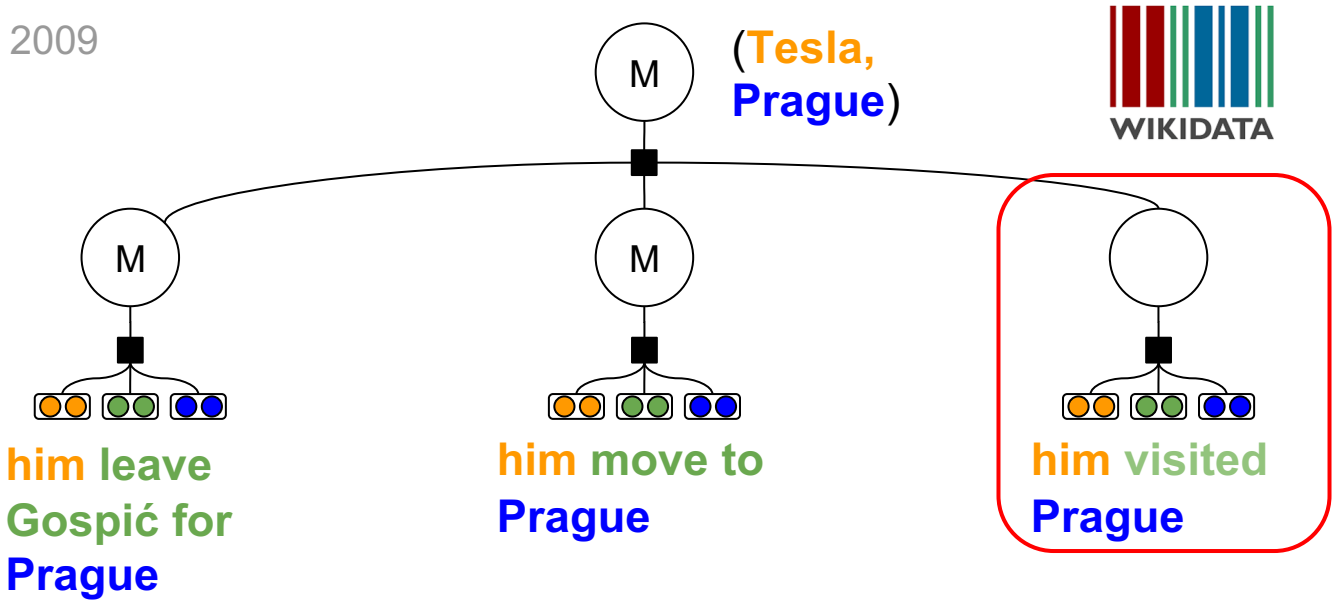
him move to Prague



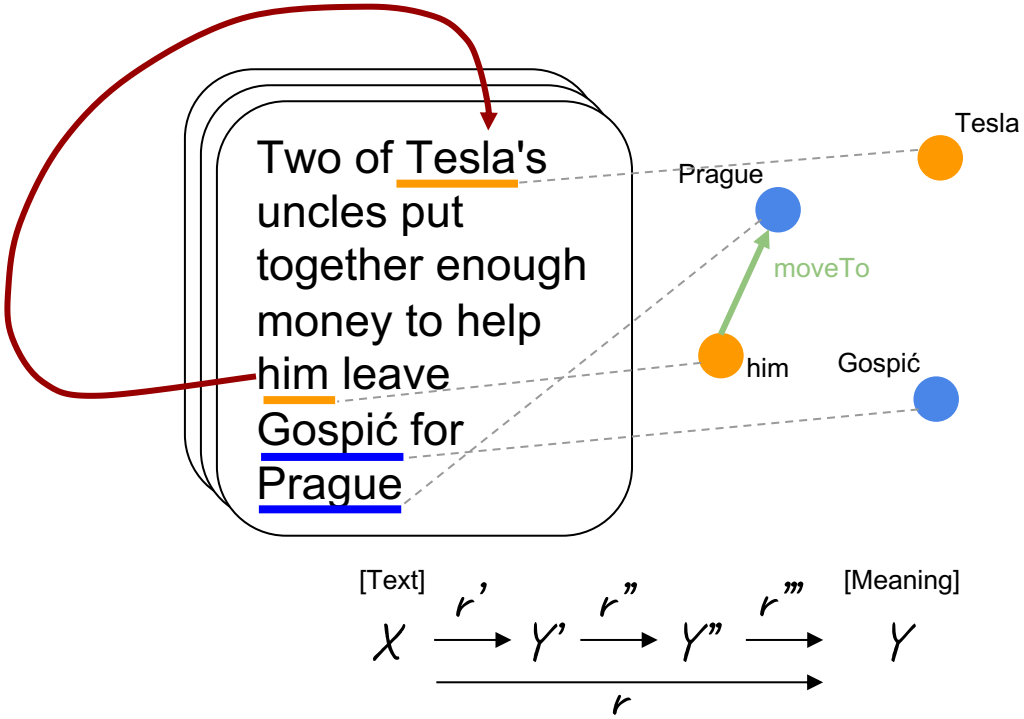
him settle in Prague

Distant Supervision & Multiple Instance Learning

Mintz et al., 2009

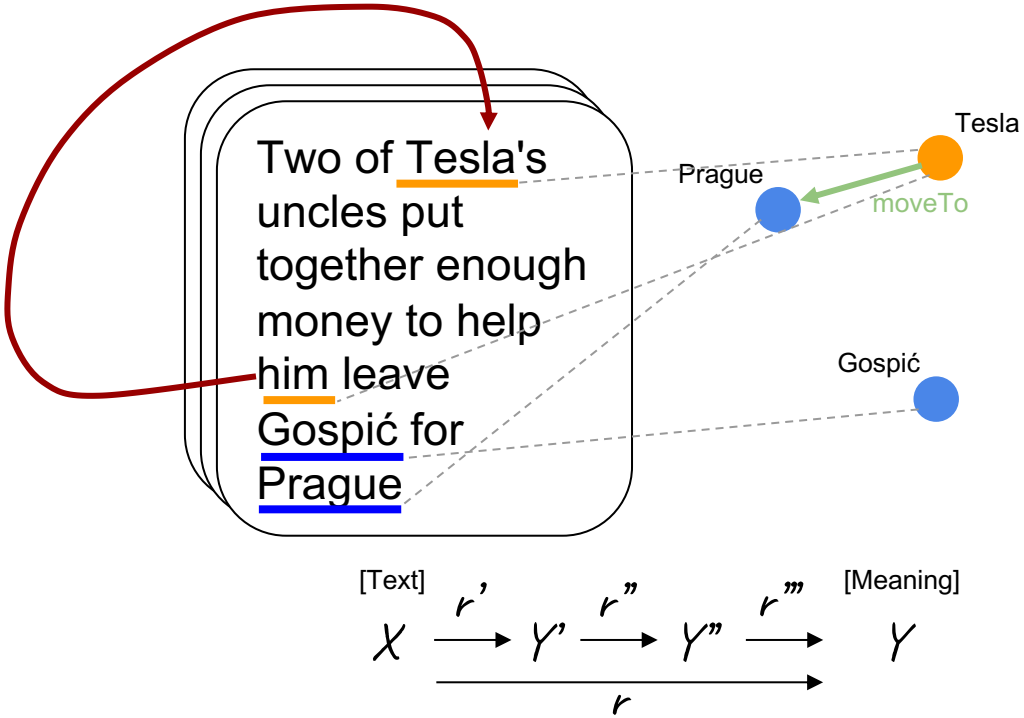


Coreference Resolution

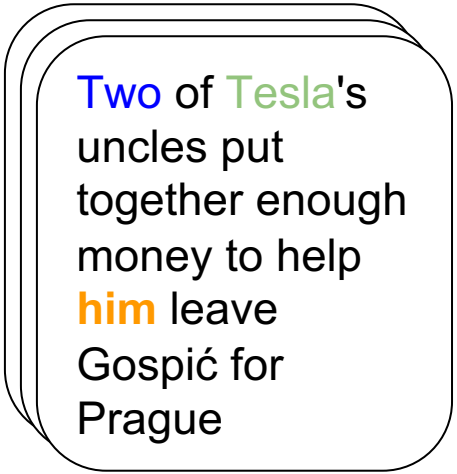


- Neural Classification
- Latent Variables

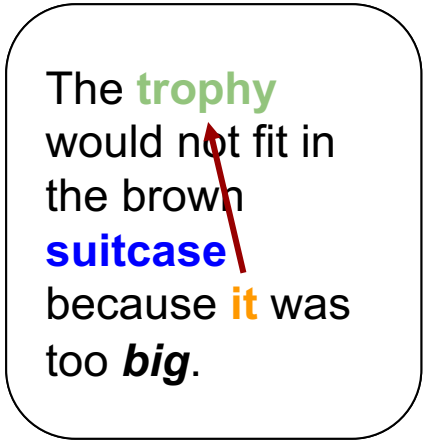
Collapsing Nodes

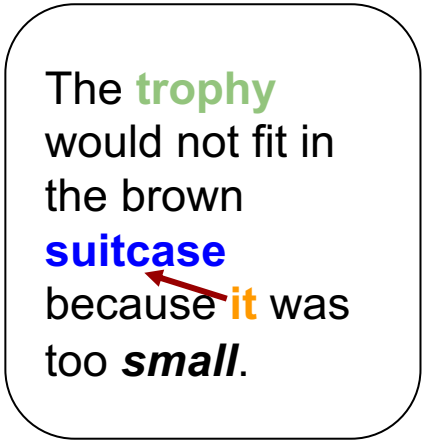


Challenge: Common Sense



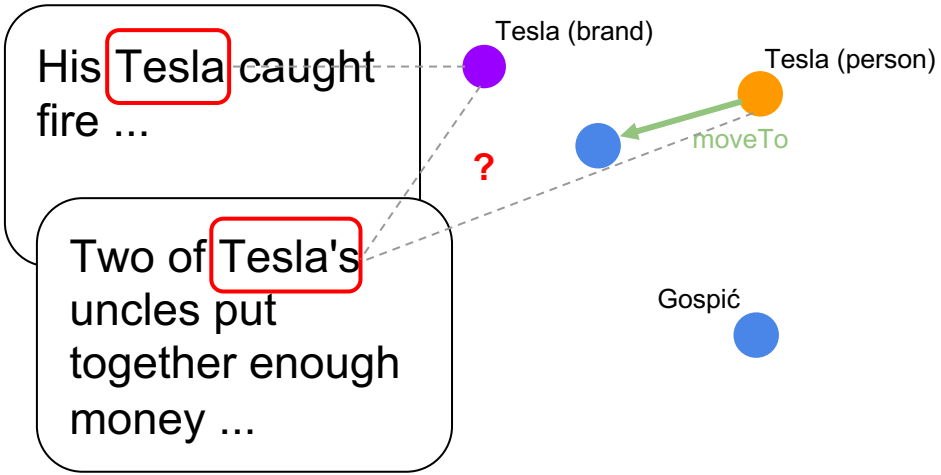
Surface





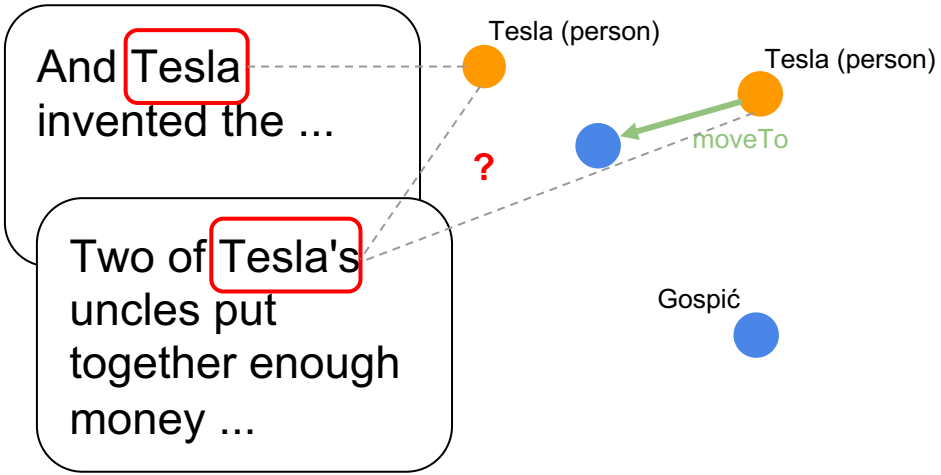
Common Sense

Entity Linking



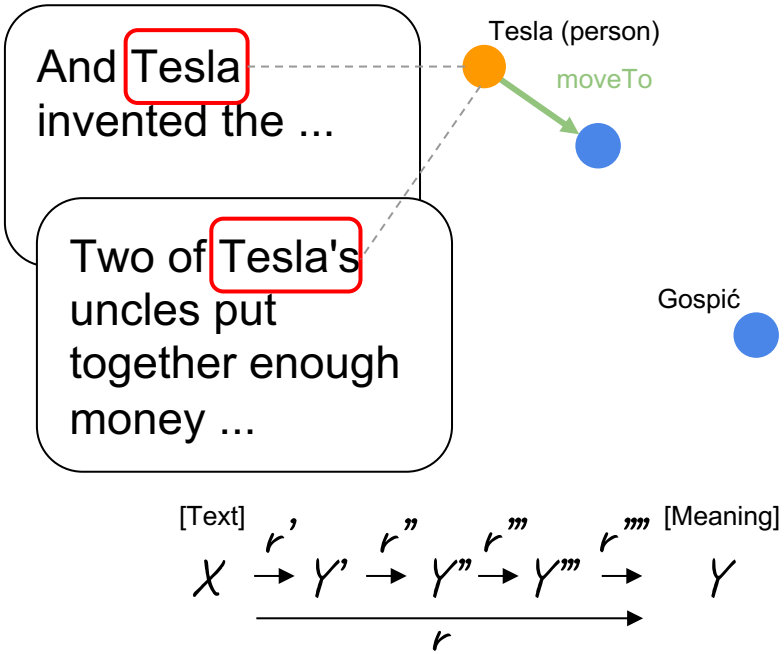
- Reranking ...
- Embeddings ...

Entity Linking

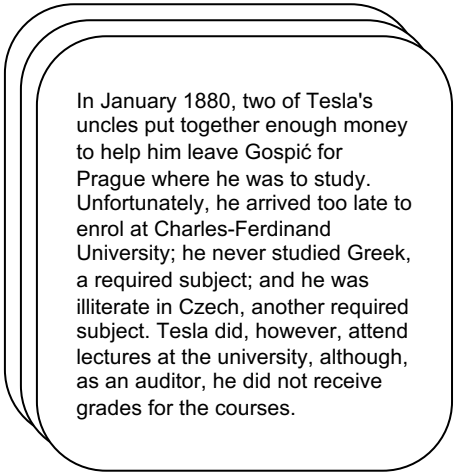


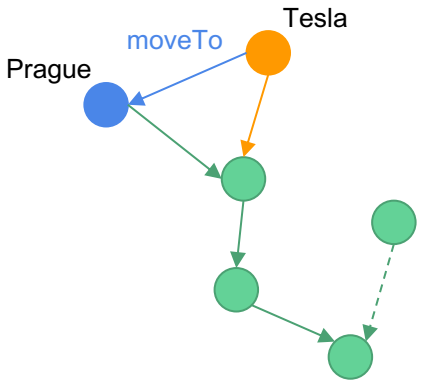
- Reranking ...
- Embeddings ...

Collapsing


And Tesla invented the ...
Two of Tesla's uncles put together enough money ...
Tesla (person)
moveTo
Gospić
[Text]
[Meaning]

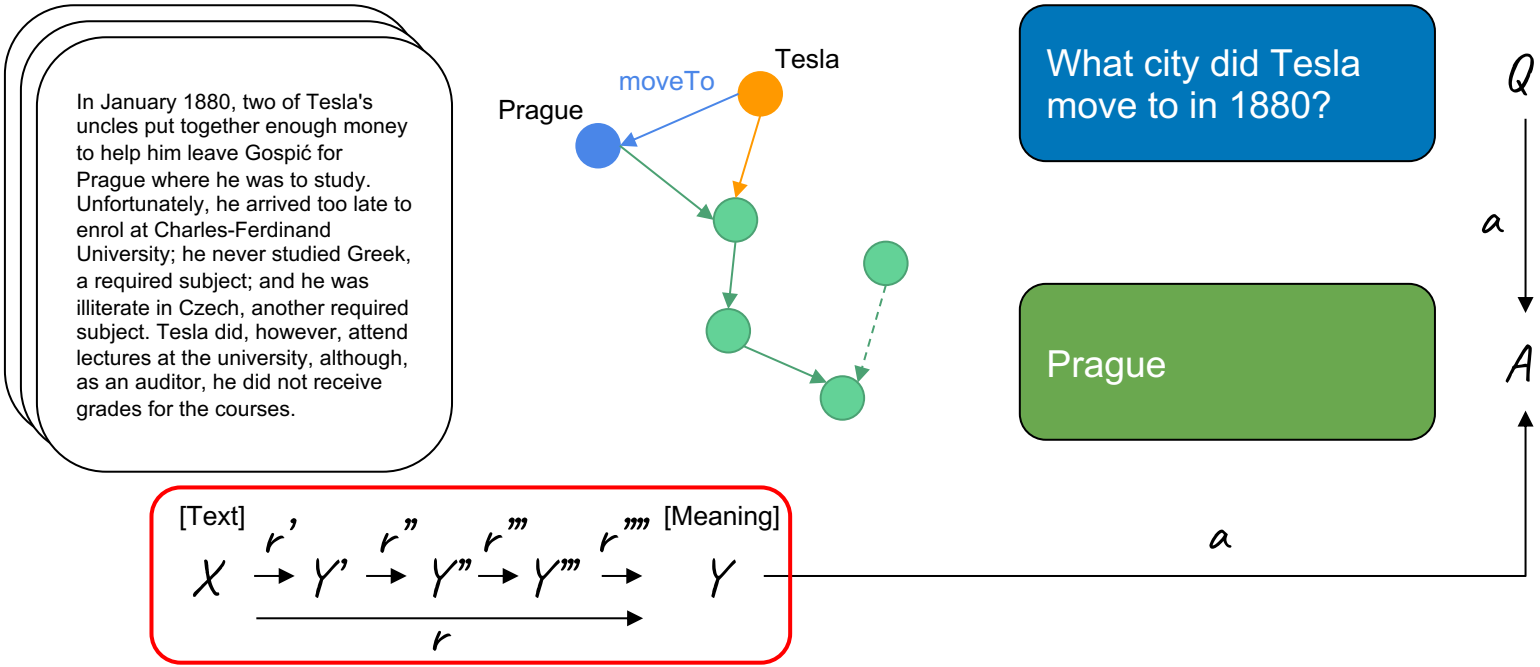
Strengths





- Supports Reasoning
- Fast access
- Generalisation
- Interpretable
- Existing KBs can serve as supervision signal!

Weakness: Cascading errors

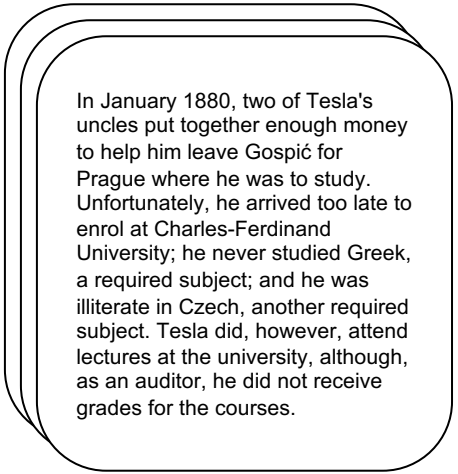


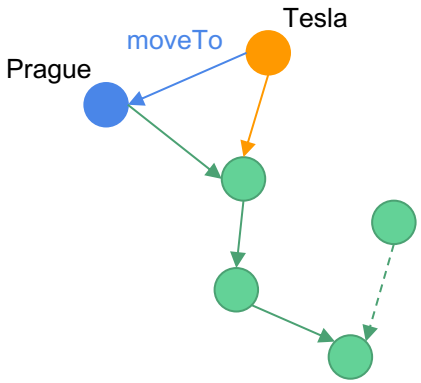
Weakness: Cascading errors

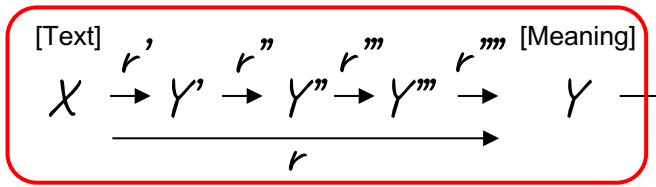
What city did Tesla move to in 1880?

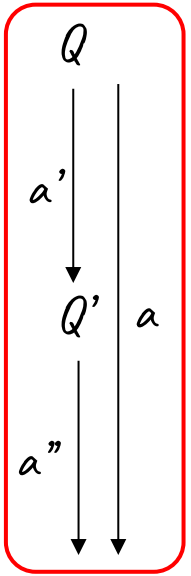
moveTo(Tesla,X)?

Prague



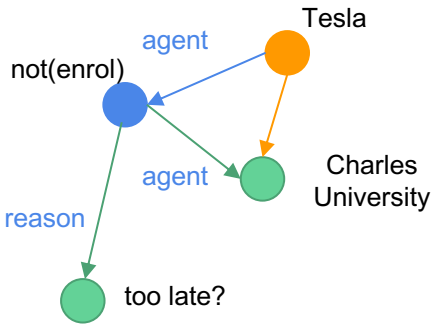






Weakness: Engineering Schemas and Formalisms

Unfortunately, he arrived too late to enrol at Charles University



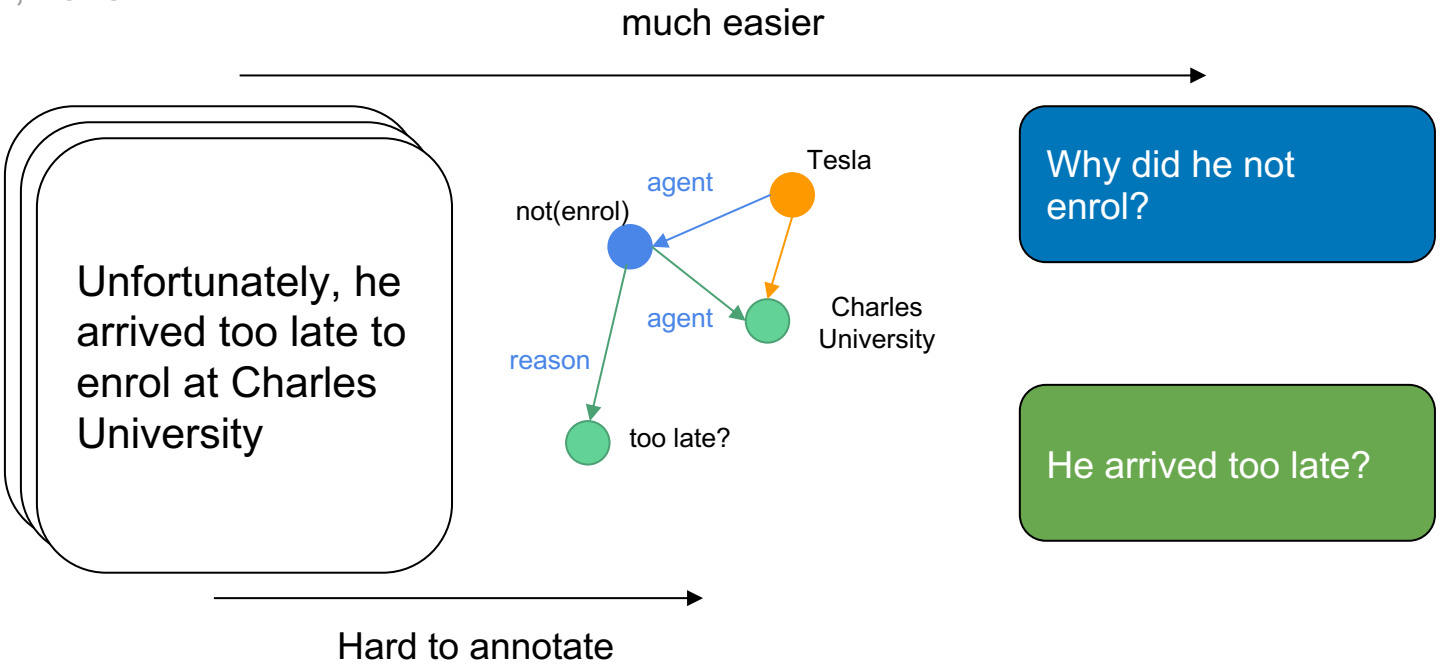
Why did he not enrol?

He arrived too late?

getting this right is hard

Weakness: Annotation

He et al., 2015

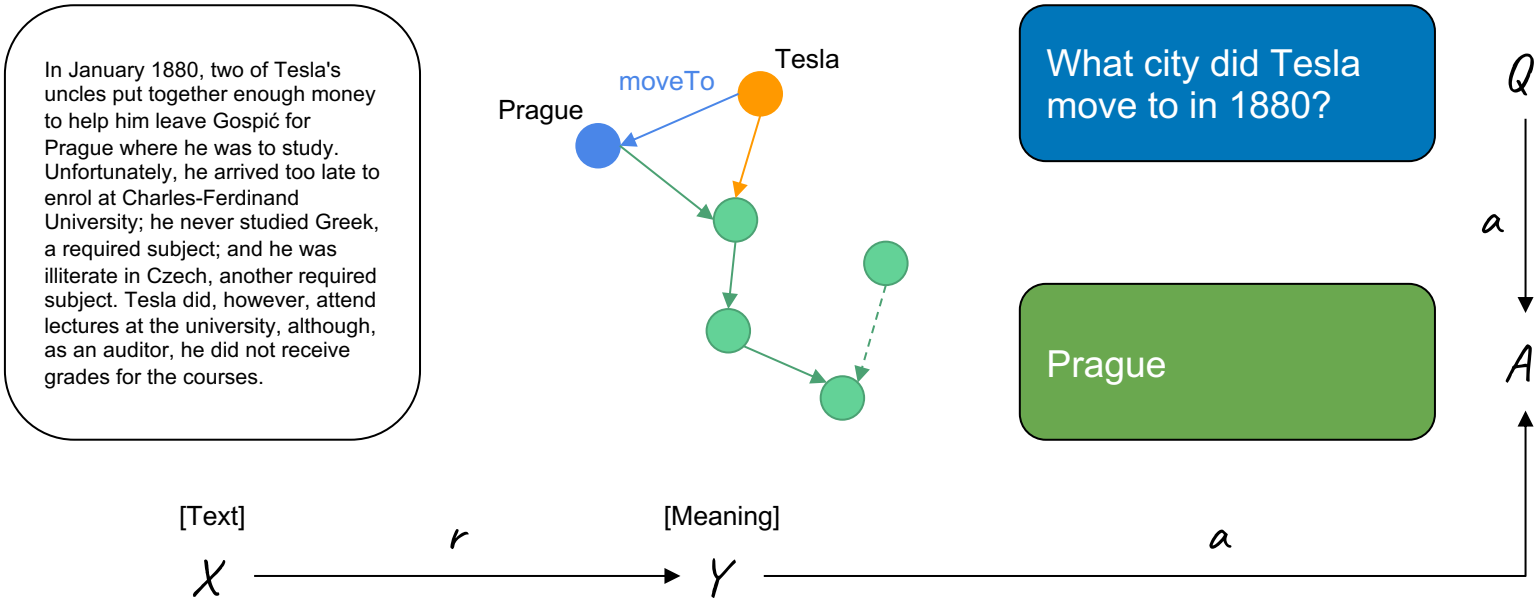


Structured Representations

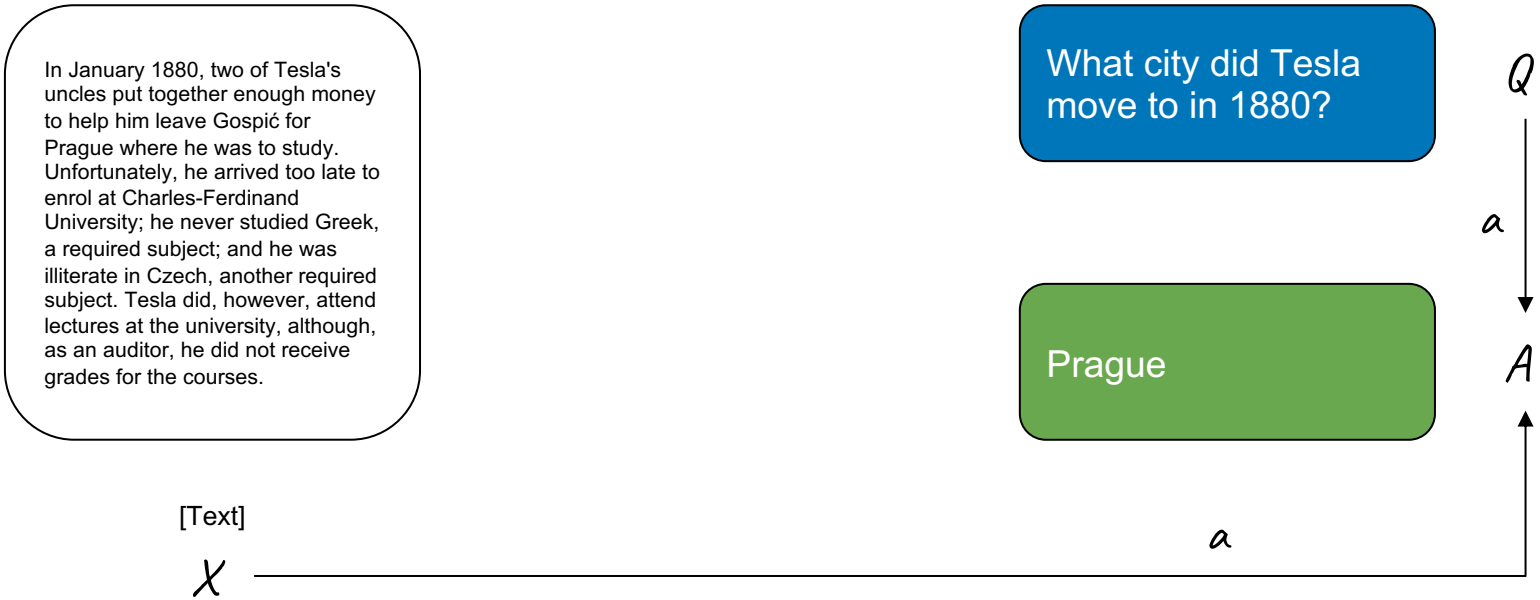
- Advantages
  - Fast access
  - Scalable
  - Interpretable
  - Supports reasoning
  - Universality of representations: independent of question

- Disadvantages
  - Less robust to variation in language
  - Cascading errors
  - Schema engineering
  - Annotation requires experts

Is there another way?



Omitting Symbolic Meaning Representations !!



Machine Reading AT SCALE

A **machine** processes **a (very) large collection of texts** to satisfy an **information need**

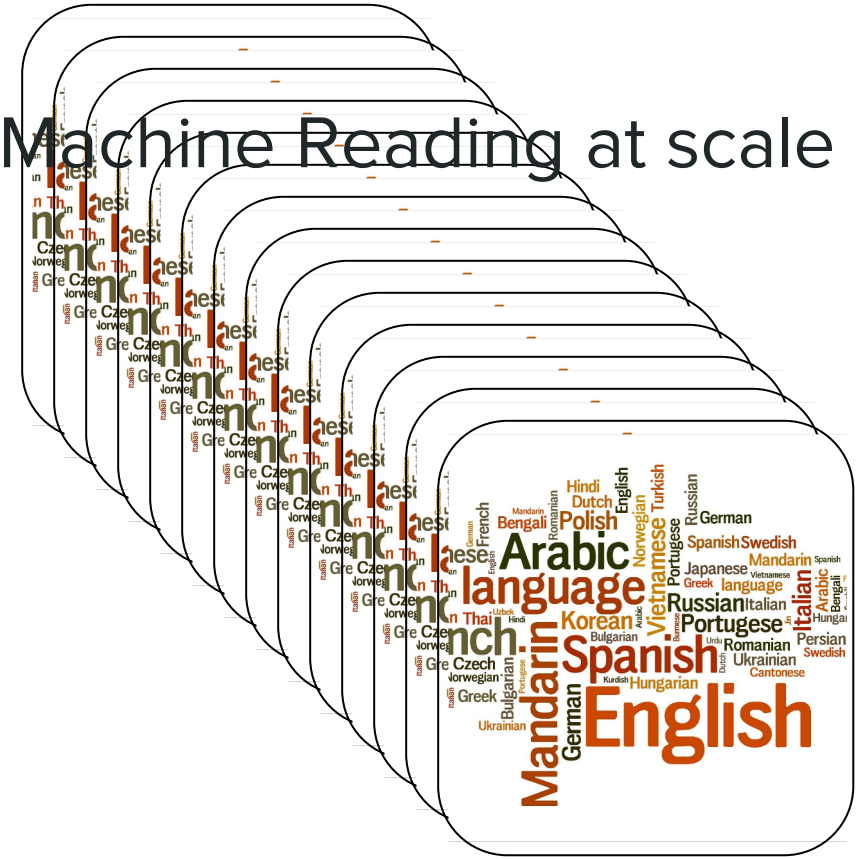
Machine Reading

[Text]

uses for

[Information Need]

Machine Reading at scale



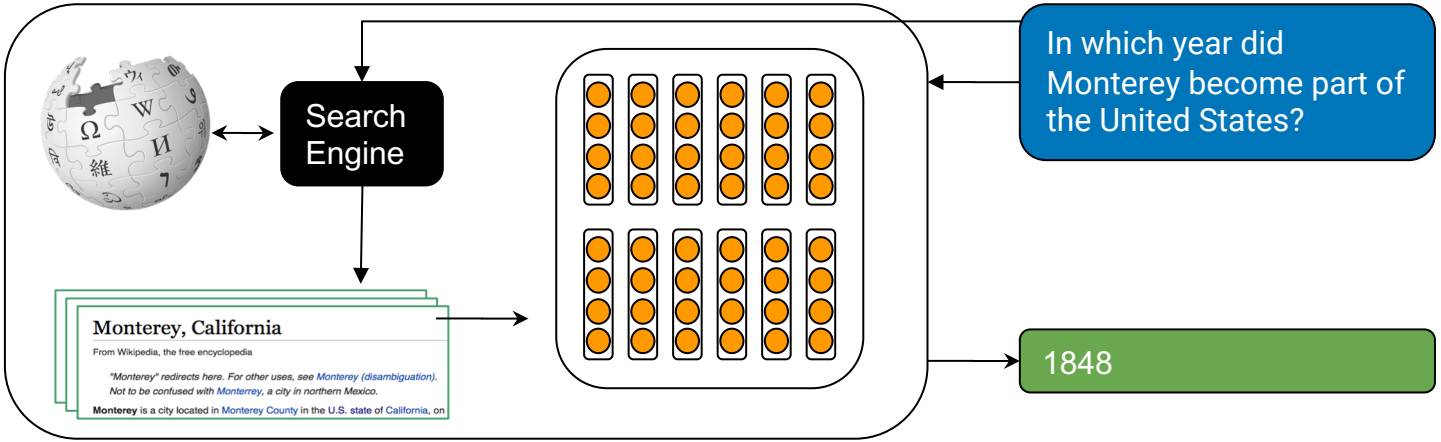
[Large collection of Text]

uses for

[Information Need]

Typical Machine Reading at Scale System

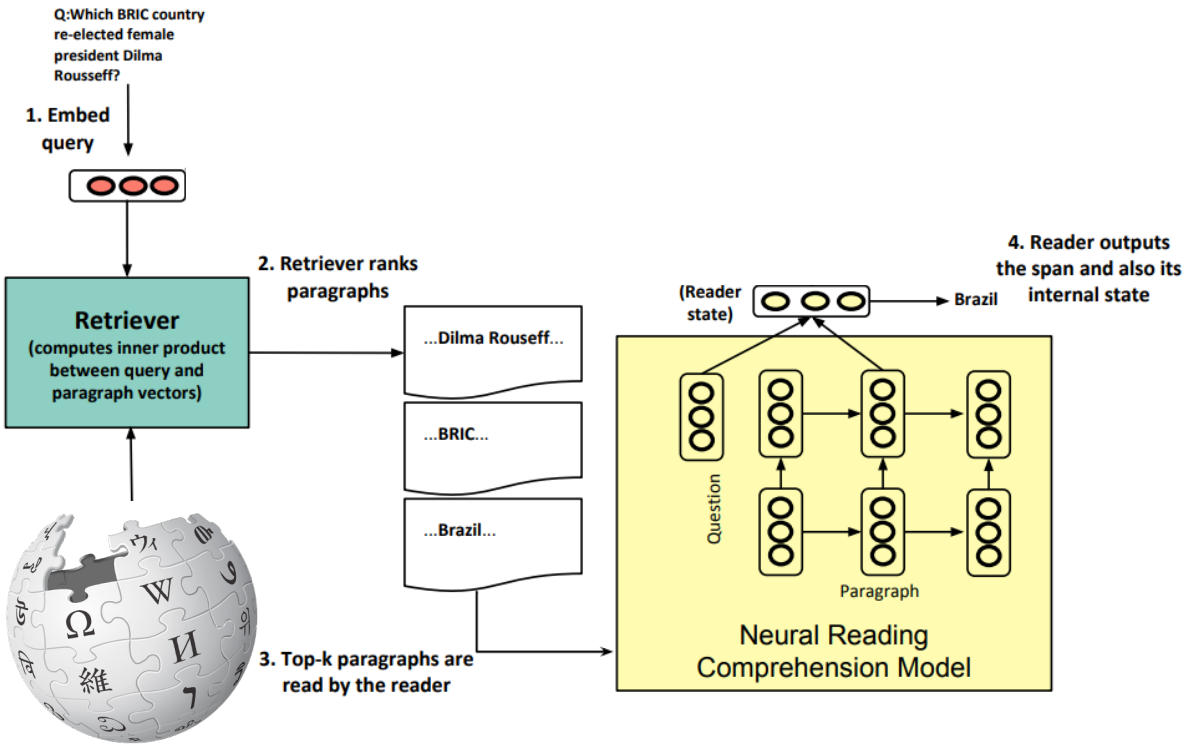
Dr.QA Chen et al., 2017



No way to recover if the search engine is wrong!

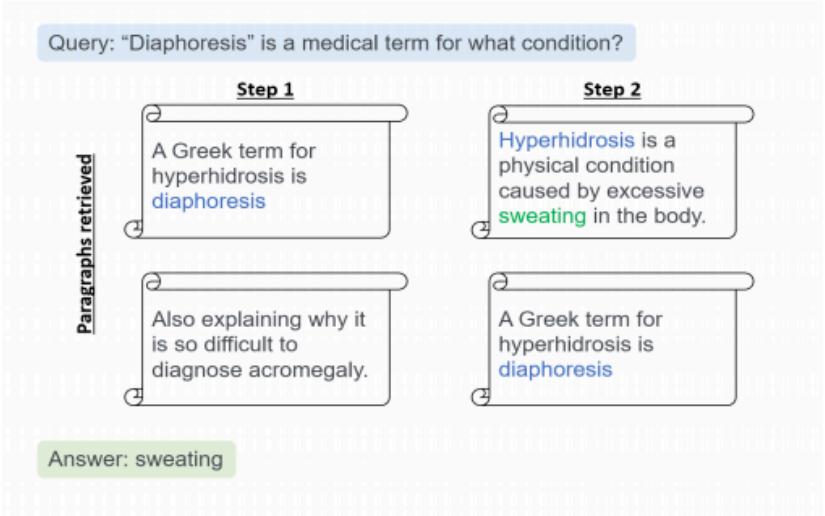
Current best: Multi-Step Retriever-Reader

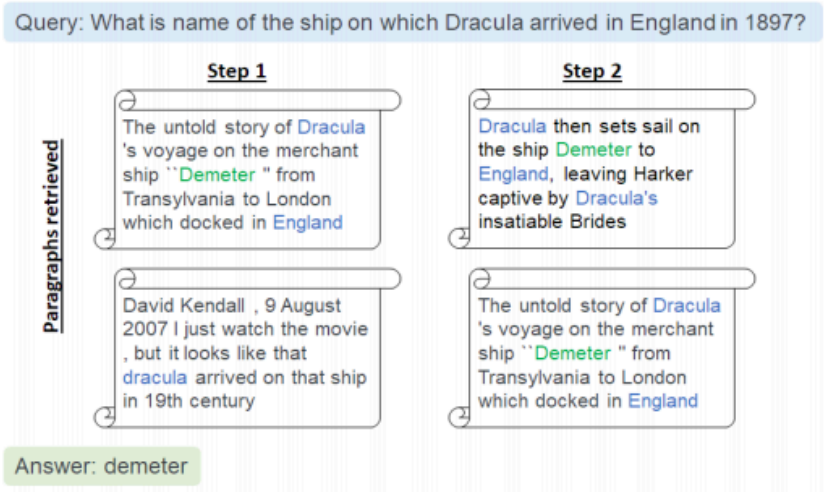
Das et al., 2019



Current best: Multi-Step Retriever-Reader

Das et al., 2019



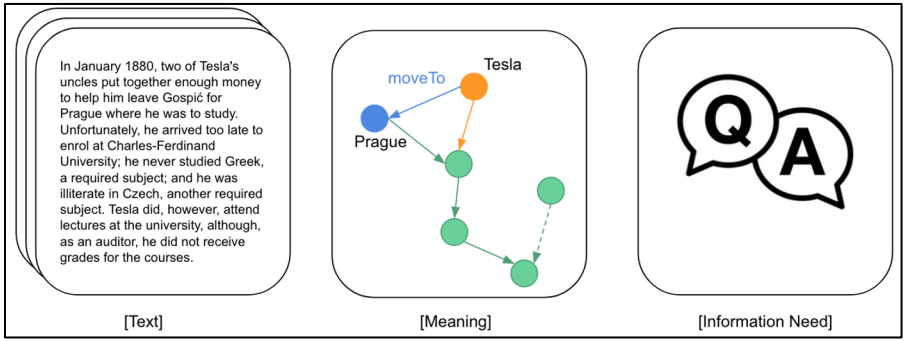


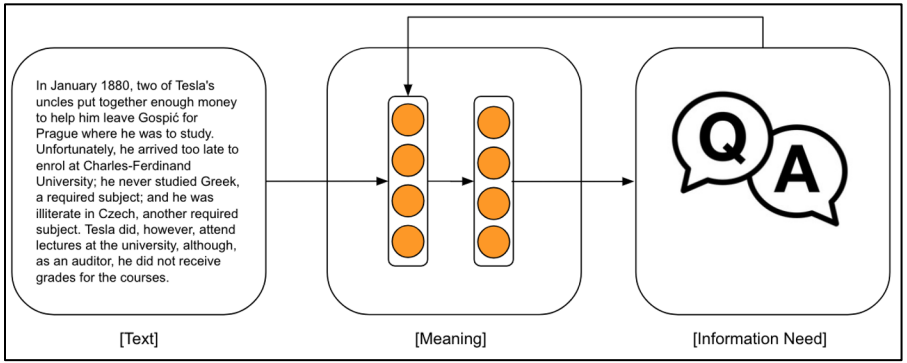
Between 40 and 60% of correct responses (for rather simple questions)

A Paradigm Shift

- Symbolic Meaning Representations
- ➡ Latent Vector Representations

- Feature Engineering & Domain Expertise
- ➡ Architecture Engineering & ML/DL Expertise





Pros and cons

End-to-end models	Symbolic systems
Neural Networks	*KBs, Inductive Logic Programming, etc.*
• Scale to very large datasets • Can be used by non domain experts • Robust to noise and ambiguity in data • Game changers in multiple applications	• Small scale conditions • Require heavy expert knowledge • Very brittle with noisy, ambiguous data • Limited applicative success
• Very data hungry (mostly supervised data) • Can't learn easily new tasks from old ones • Not interpretable • Relatively simple reasoning	Great research opportunities! ostractions

Current Challenge: Reconciling Conflicting Information

So how much does the UK pay to the EU per week?

"Once we have settled our accounts, we will take back control of roughly **£350m** per week." *Boris Johnson*

"We are not giving £20bn a year or £350m a week to Brussels - Britain pays **£276m** a week to the EU budget because of the rebate." *BBC Reality Check*

"...When those are taken into account the figure is **£250m**." *Independent*

?

Trust into source, timeline, ...

Conclusion

- We've seen 2 approaches for building system to answer any question
- Most deployed systems still rely on traditional pipelines for the most part (+ some DL here and there)

- Why? Scale, reliability, interpretability

- Open questions:
  - All shortcomings of Machine Reading → Open domain QA. Need to solve them
  - Will pretrained contextual embeddings change everything forever?
  - Can we combine both symbolic and end-to-end approaches?